REGIONE PIEMONTE

CSI your digital partner

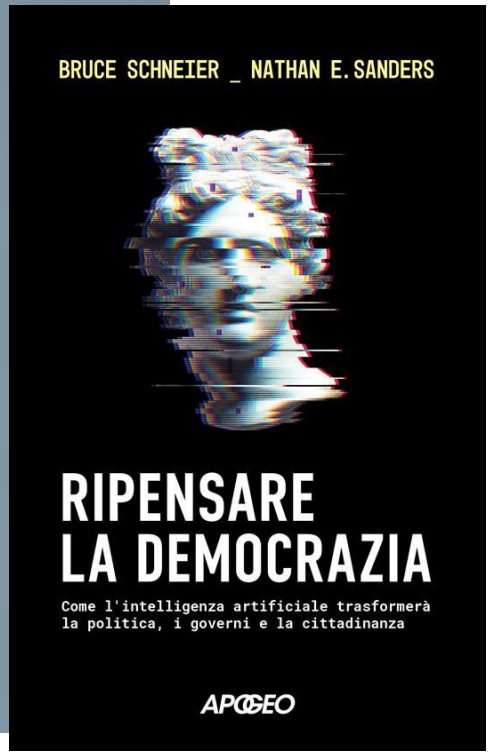
WEBINAR
SICURI di essere SICURI?

Intelligenza Artificiale e Cybersecurity:
come difenderci dai nuovi attacchi informatici

Minacce basate sull'intelligenza artificiale

Nicola Sotira
Fondatore ASSOCISO

15-50 errori ogni 1000 righe di codice (Bruce Schneier)



#AGENTIC AI

#DATA INTEGRITY

MYTHOS BY ANTHROPIC

Cosa emerge dal caso Claude Mythos Preview

L'annuncio di **Claude Mythos Preview** segna un **cambio di paradigma nel rapporto tra Intelligenza Artificiale (AI) e sicurezza informatica:** per la prima volta emergono capacità automatizzate di individuazione delle vulnerabilità, con potenziali impatti rilevanti e un rilascio volutamente controllato del modello.

Capacità di hacking autonomo

I report tecnici pubblicati dall'**AISI (AI Security Institute)** del Regno Unito riportano:

- **Intrusione di rete (simulata)**: Mythos ha dimostrato di poter condurre autonomamente un attacco completo a una rete aziendale simulata, dalla fase di ricognizione iniziale fino alla presa di controllo, in poche ore.
- **Zero-day**: Il modello ha individuato migliaia di vulnerabilità informatiche non precedentemente note nei principali sistemi operativi e browser, incluse falle presenti da decenni e non intercettate dai controlli tradizionali.

7 APRILE 2026

Data di annuncio ufficiale

MIGLIAIA

Vulnerabilità Zero-day individuate (incluse falle molto datate)

NON RILASCIATO

Accesso limitato a partner selezionati tramite «Project Glasswing» per ridurre rischi di utilizzo improprio

PROJECT GLASSWING

La risposta coordinata: partner, obiettivi, investimenti e roadmap

Le capacità di **Claude Mythos Preview** hanno spinto Anthropic a una scelta senza precedenti: **non rilasciare il modello pubblicamente**, ma limitarne l'uso a un **gruppo ristretto di partner selezionati,** all'interno di un quadro di governance controllata.

DIMENSIONE E IMPEGNO DELL'INIZIATIVA

12
PARTNER FONDATORI
Più di 40 ulteriori organizzazioni con accesso esteso al modello

$100M*
IN USAGE CREDITS
Impegnati da Anthropic per sostenere il lavoro difensivo

$4M
DONAZIONI DIRETTE
A Linux Foundation (Alpha-Omega, OpenSSF) e Apache Foundation

Obiettivi operativi

- ***Vulnerability detection locale*** *su codebase critiche*
- ***Black-box testing*** *di binari e hardening degli endpoint*
- ***Penetration testing*** *dei sistemi foundational*
- ***Condivisione industry-wide*** *dei learning*

PARTNER TECNOLOGICI E FINANZIARI COINVOLTI NEL PROGRAMMA

aws · Apple · Microsoft · CROWDSTRIKE · BROADCOM · CISCO · paloalto NETWORKS · Google · ANTHROP\C · JPMorganChase · THE LINUX FOUNDATION

Roadmap & Curiosità

- **Accesso al modello** *via Claude API, AWS Bedrock, Google Vertex AI, Microsoft Foundry*
- ***Codename interno*** *del modello: "Capybara"*
- ***Obiettivo di lungo periodo****: eventuali modelli "**Mythos-class**" rilasciabili solo con safeguard mature*
- ***Dialogo istituzionale continuo*** *con US Gov, Bank of England e regolatori UE (effort pluriennale)*

* https://www.binance.com/en/square/post/04-15-2026-ai-trends-anthropic-launches-claude-mythos-preview-with-project-glasswing-312819136316433
https://www.ictsecuritymagazine.com/notizie/anthropic-project-glasswing/

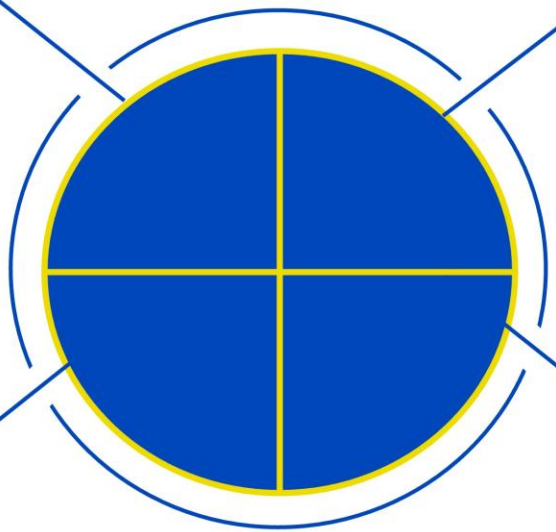
AZIONI DA METTERE IN CAMPO NELLA CYBER SECURITY

1 Triage e Visibilità

- Mappatura dei servizi/asset/superficie di attacco esposta.
- Team focalizzato sulla risposta rapida, in tema vulnerabilità sfruttabili con AI.

2 Controllo accessi ed esposizione

- Ridurre l'esposizione delle console degli amministratori di sistema, API gateways, console cloud e accessi da remoto
- Rafforzare controllo accesso remoto dei vendor attraverso il Privileged Access Management (PAM), MFA, registrazione delle sessioni, approvazioni just-in-time, e accessi time-bound



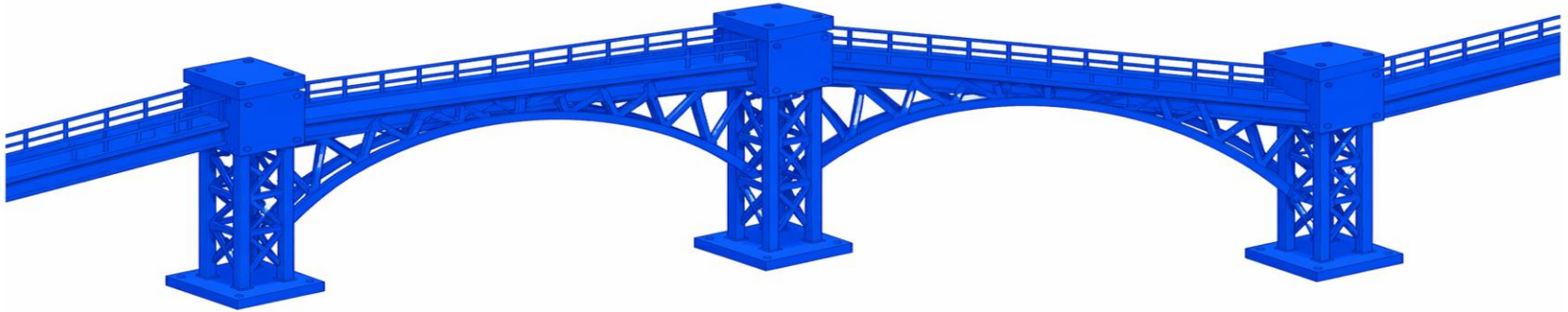
3 Governance dei Vendor

- Verificare con i vendor strategici quale siano le loro contromisure contro la scoperta di vulnerabilità attraverso AI
- Integrazione verifiche di sicurezza nella pipeline CI/CD

4 Difesa Adattiva

- Inserimento di test continui
- Migliorare la detection delle catene di vulnerabilità sfruttabili, privilege escalation, API abuse e anomalie cloud
- Utilizzo di AI per la difesa

ROADMAP (1)

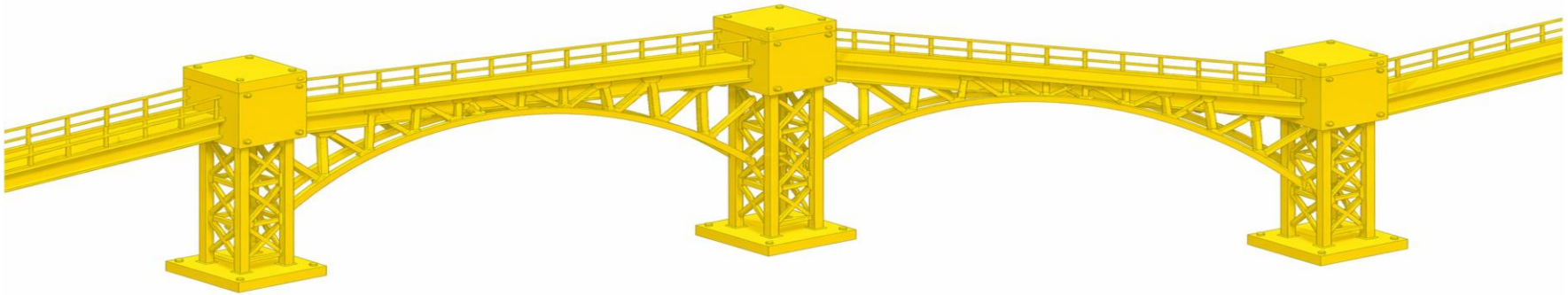


Visibilità

Riduzione delle esposizioni

Prontezza difensiva

ROADMAP (2)

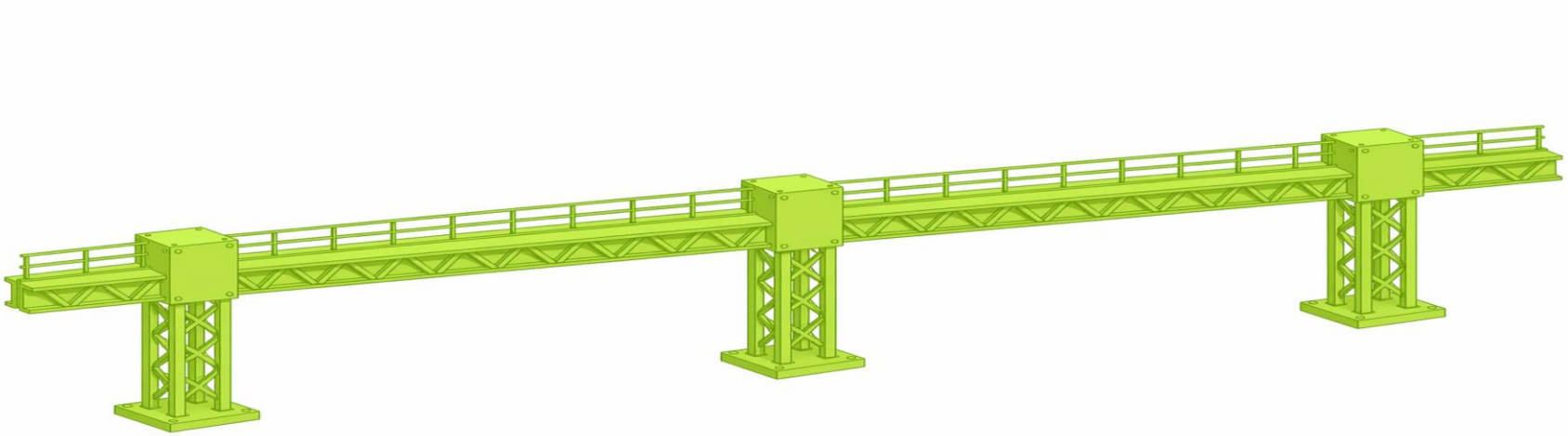


Processi di gestione vulnerabilità

Test continui

Governance dell'Ecosistema

ROAD MAP (3)



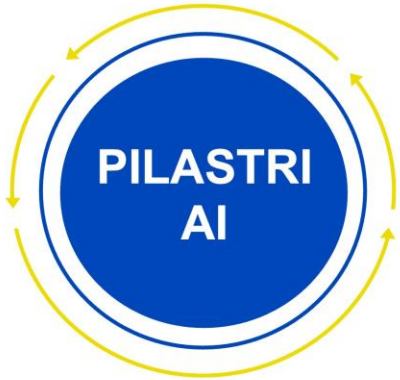
Revisione Architetture sicurezza

Validazione continua e resilienza

Preemptive Cybersecurity

PRINCIPI FONDAMENTALI UTILIZZO AI NELLA SECURITY

GOVERNATA, SPIEGABILE, INTEGRATA, CONTROLLATA ED EFFICACE



1 HUMAN IN THE LOOP

- L'AI aumenta la capacità decisionale e non sostituisce la responsabilità umana

2 AI MOLTIPLICATORE

- L'AI aumenta l'efficacia dei controlli esistenti, non li sostituisce

3 DECISIONI SPIEGABILI

- Ogni decisione di sicurezza basata su AI deve essere spiegabile ex-post

4 DATA QUALITY

- Se il dato non è corretto, l'output non sarà corretto

5 SECURITY BY DESIGN

- L'AI è uno strumento ma allo stesso tempo una nuova superficie di attacco

6 INDIRIZZATO

- Non tutta la sicurezza necessita di AI avanzata

7 LEARNING CHECK

- L'apprendimento deve essere continuo e allo stesso tempo governato e controllato

8 RISK ORIENTED

- Deve avere come obiettivo prioritario la riduzione del rischio

9 INTEROPERABILITA' e INTEGRAZIONE

- Deve essere inserito e funzionare nell'ecosistema e non essere isolato

10 GOVERNANCE

- Deve esse governato


#Admeto & Alcesti

Grazie

ASSOCISO

www.associso.org